Università di Pisa

Topics for Projects

Giuseppe Attardi

Human Language Technologies

Dipartimento di Informatica

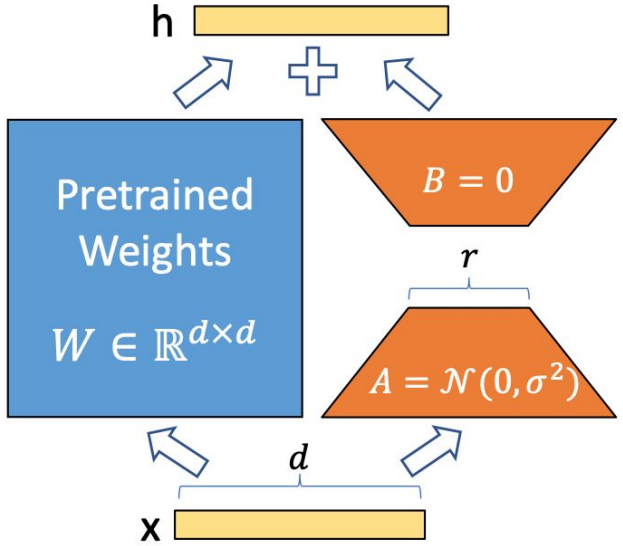
Università di Pisa

Project Types

- **Task oriented**. Find an application/task of interest and explore how to approach/solve it effectively, often with an existing model
- **New architecture**. Implement a new or complex neural architecture and demonstrate its effectiveness
- **Analysis project**. Analyze the behavior of a model: how it represents linguistic knowledge or what kinds of phenomena it can handle or errors that it makes
- **Theoretical project**. Show some interesting, non-trivial properties of a model type, data, or a data representation

Question Generation

- INVALSI (https://www.invalsi.it/invalsi/)
  - Question generation from Wikipedia articles
  - Fine tune a LLM on generating questions from articles:
    - Task 1: generate question (e.g. Who won the Super Bowl 50?) given subject (e.g. Denver Broncos)
    - Task 2: extract subject (e.g. Denver Broncos) from article (e.g. Super Bowl 50 was an American football game to determine the champion of the National Football League (NFL) for the 2015 season. …)
  - Train using a QA dataset used in reverse, i.e. to generate a question given the answer and a source text.
  - For example, use SQuAD as a training set https://towardsdatascience.com/nlp-building-a-question-answering-model-ed0529a68c54
  - Try using Alpaca-LoRA https://github.com/tloen/alpaca-lora
- Alternatively, use PAQ:
  Alberti et al. (2019) Synthetic QA corpora generation with roundtrip consistency. *arXiv preprint arXiv:1906.05416*



QA with RLHF

- Use the StackExchange dataset (a subset of 10-100,000 pairs) to train LlaMA with Reinforcement Learning for asnwering questions
- https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences
- Reinforcement Learning with Human Feedback
  1. Generate responses from prompts
  2. Rate the responses with the reward model
  3. Run a reinforcement learning policy-optimization step with the ratings
- Ask me for the 7B LlaMA model converted to HF format

Legal Judgement Predictor

- Dataset from a database of the European Court of Human Rights (ECHR)
- Chalkidis et al. (2019). Neural legal judgment prediction in English.
- https://aclanthology.org/P19-1424.pdf

Challenges

- Key Point Analysis
  - https://github.com/ibm/KPA_2021_shared_task
- Language Generation from Structured Data
  - https://webnlg-challenge.loria.fr/workshop_2020/
- COVID-19 Global Hackathon
  - https://covid-global-hackathon.devpost.com/
- BioASQ (http://bioasq.org)
- CLEF CheckThat! Lab (https://gitlab.com/checkthat_lab/clef2021-checkthat-lab)
- Pharma CoNER (http://temu.bsc.es/pharmaconer/)
- Loop Q Prize (https://www.loopqprize.ai)
- The Conversational Intelligence Challenge 2 (http://convai.io)

Key Point Analysis

New NLP Task: Given an input corpus, consisting of a collection of relatively short, opinionated texts focused on a topic of interest, the goal of KPA is to produce a succinct list of the most prominent key-points in the input corpus, along with their relative prevalence. Thus, the output of KPA is a bullet-like summary, with an important quantitative angle and an associated well-defined evaluation framework.

Applications: to gain better insights from public opinions as expressed in social media, surveys, parliamentary debates, etc.

Challenge: https://github.com/ibm/KPA_2021_shared_task

Tracks:

1. Key-Point Matching
2. Key Points Generation and Matching

KPA Example

Obtained by human labeling on key points provided by an expert, on the topic "Homeschooling should be banned", on the pro stance arguments.

Key point	Matched arguments count
Mainstream schools are essential to develop social skills.	61
Parents are not qualified as teachers.	20
Homeschools cannot be regulated/standardized.	15
Mainstream schools are of higher educational quality.	9

KPA Example

Argument	Matching key point
children can not learn to interact with their peers when taught at home	Mainstream schools are essential to develop social skills
homeschooling a child denies them valuable lifeskills, particularly interaction with their own age group and all experiences stemming from this.	
to homeschool is in one way giving a child an immersive educational experience, but not giving them the social skills and cooperative skills they need throughout life, so should be banned.	
parents are usually not qualified to provide a suitable curriculum for their children. additionally, children are not exposed to the real world.	Parents are not qualified as teachers
it is impossible to ensure that homeschooled children are being taught properly	Homeschools cannot be regulated/standardized.

Track 1 - Key-Point Matching

Input:

Arguments and expert key points for topic and stance in the test dataset. The input consist of three files:

- arguments.csv - lists all the arguments for each topic, along with the stance of each argument towards the topic.
- key_points.csv - lists all the key points for each topic, along with the stance of each key point towards the topic.
- labels.csv - labeled pairs of argument id and key point id. Note that not all the possible pairs are labeled.

Output :

For each argument, its match score for each of the key points under the same topic and in the same stance towards the topic.

{"arg_15_0": {"kp_15_0": 0.8282181024551392, "kp_15_2": 0.9438725709915161},
"arg_15_1": {"kp_15_0": 0.9994438290596008, "kp_15_2":0}}

Text Generation from Structured Data

- Big language models generate readable sentences, but often sematically incorrect
- Example (BlenderBot, Transformer based):

Who is the Italian prime minister?

[BlenderBot]: **It is Mussolini. He was the leader of the fascist party during WWII.**

Is he still alive?

[BlenderBot]: **Yes, he is still alive. He has been in politics since 1989. He also has a son.**

Text Generation from Structured Data

Example (WebNLG)

D= Aarhus | leaderName | Jacob Bundsgaard
Linearized D= <subject> Aarhus **<predicate>** leader name **<object>** Jacob Bundsgaard

T= The leader of Aarhus is Jacob Bundsgaard.

- Challenge: https://webnlg-challenge.loria.fr/workshop_2020/
- Code: https://github.com/amazon-research/datatuner

Prompt Tuning

- Keep the LM frozen and learn “soft prompts” to condition frozen language models to perform specific downstream tasks
- https://github.com/google-research/prompt-tuning
- Data sets:
  - CrossFit: https://github.com/INK-USC/CrossFit

BERTology

- Exploring which linguistic knowledge is incorporated in transformers that enables them to perform other tasks than LM
- Probes: e.g. syntax probes

Chatbot

- Alexa Prize
  - https://developer.amazon.com/alexaprize
- Alexa Topical Chat Dataset
  - https://github.com/alexa/alexa-prize-topical-chat-dataset
  - Identify transitions between topics
  - Suggest sources of information
- The Conversational Intelligence Challenge 2 (ConvAI2)
  convai.io/
- ACCENTOR: Adding Chit-Chat to Enhance Task-Oriented Dialogues
  - https://github.com/facebookresearch/accentor

IWPT Shared Task

- The Enhanced Universal Dependency Shared Task at IWPT 2021 involves dependency parsing from plain text.
- This involves several subtasks:
  - Tokenization using DL
  - POS using DL
  - Morphological analysis
  - Dependency parsing
  - Enhanced dependencies

CoNLL 2018: Deep Learning Tokenizer

- CoNLL 2018 challenge requires a tokenizer for all the Universal Dependency TreeBanks
- Build a MultiWord Tokenizer similar to Stanza, but exploiting clusters of embeddings as POS surrogates (cooperation with Stanford) https://stanfordnlp.github.io/stanza/mwt.html

Evalita 2016-2023

- www.evalita.it/2016
  - POSTWITA
  - QA4FAQ
  - NEEL-IT
- www.evalita.it/2018
  - ABSITA
  - HaSpeeDe
  - NLP4FUN (more statistics than linguistics?)
- http://www.evalita.it/2020
  - Affect, Hate, and Stance

Evalita 2023

- https://www.evalita.it/campaigns/evalita-2023/tasks/
- EMit – Categorical Emotion Detection in Italian Social Media
- EmotivITA – Dimensional and Multi-dimensional emotion analysis
- PoliticIT – Political Ideology Detection in Italian Texts
- GeoLingIt – Geolocation of Linguistic Variation in Italy
- LangLearn – Language Learning Development
- HaSpeeDe 3 – Political and Religious Hate Speech Detection
- MULTI-Fake-DetectiVE – MULTImodal Fake News Detection and Verification
- ACTI – Automatic Conspiracy Theory Identification

Evalita 2023

- NERMuD -Named-Entities Recognition on Multi-Domain Documents
- CLinkaRT – Linking a Lab Result to its Test Event in the Clinical Domain
- WiC-ITA – Word-in-Context task for Italian
- DisCoTEX – Assessing DIScourse COherence in Italian TEXts

Question Answering Tasks

- Tensorflow 2.0 QA
  - https://www.kaggle.com/c/tensorflow2-question-answering
- SemEval 2017
  Task 3
- Evalita 2016
  QA4FAQ
- SQuAD
  https://towardsdatascience.com/nlp-building-a-question-answering-model-ed0529a68c54
- Movie QA
  http://movieqa.cs.toronto.edu/home/
- StackExchange dataset
  https://huggingface.co/datasets/HuggingFaceH4/stack-exchange-preferences
- Natural Language Interfaces for Web of Data (NLIWoD4)
  http://2018.nliwod.org/challenge

Chatbots

- AWS Chatbot Challenge
  - https://aws.amazon.com/events/chatbot-challenge/
- Ubuntu Dialog Corpus:
  - https://github.com/rkadlec/ubuntu-ranking-dataset-creator

Neural Machine Translation

- English-Italian
  - Europarl Corpus or http://www.manythings.org/anki/
  - https://www.tensorflow.org/addons/tutorials/networks_seq2seq_nmt
- References:
  - D. Bahdanau, K. Cho, Y. Bengio. Neural machine translation by jointly learning to align and translate. http://arxiv.org/pdf/1409.0473v6
  - Zhang, X., & LeCun, Y. (2015). Text Understanding from Scratch. http://arxiv.org/abs/1502.01710

Twitter

- Modeling Political Bias
  - Use Italian Tweets collection
- Detecting Toxic Comments
  - Use Italian Tweets collection and Evalita 2018 HaSpeeDe corpus

Deep Learning for Sentiment Analysis

- Annotated Data: SemEval training set
  - http://alt.qcri.org/semeval2017/task4/index.php?id=data-and-tools
- Unannotated Data: 50 million tweets
- BiLSTM approach:
  - Baziotis et al. DataStories at SemEval-2017 Task 4: Deep LSTM with Attention for Message-level and Topic-based Sentiment Analysis. www.aclweb.org/anthology/S17-2126
  - Code: https://github.com/cbaziotis/datastories-semeval2017-task4

Medical texts

- Predicting side effects of drugs
  - Using collection of Italian medical record on kidney and heart diseases
- Negation/Speculative Scope Detection
  - BioScope Corpus: http://rgai.inf.u-szeged.hu/index.php?page=bioscope
- Semantic QA on medical texts:
  - BioASQ datasets: bioasq.org/

Negation/Speculation Scope

- Determine the scope of negative or speculative statements:
  - The lyso-platelet had no effect
  - MnlI-AluI could suppress the basal-level activity
- Approach:
  - Classifier for identifying cues
  - Classifier to determine scope
- Data
  - BioScope collection

Relation Extraction

- Exploit word embeddings as features + extra hand-coded features
- SemEval 2014 Relation Extraction dataset

Fake News Detection

- CLEF CheckThat! Lab
  - https://gitlab.com/checkthat_lab/clef2021-checkthat-lab
  - T1. Check-worthiness estimation
  - T2. Claim retrieval
  - T3. Fake news detection
- Stance Detection dataset for FNC-1
  - http://www.fakenewschallenge.org

Dataset Collections

- https://www.kaggle.com/datasets?search=text
- https://www.paperswithcode.com/datasets?mod=texts
- https://huggingface.co/datasets
- https://machinelearningmastery.com/datasets-natural-language-processing/
- https://github.com/niderhoff/nlp-datasets
- https://gluebenchmark.com/tasks
- https://nlp.stanford.edu/sentiment/
- https://research.fb.com/downloads/babi/